Modelos Híbridos para Reconhecimento de Emoções Faciais em Crianças

Rafael Zimmer[1], **Marcos Sobral**[2], **Helio Azevedo**[3]

rafael.zimmer@usp.br, marcossobraldev@gmail.com, hazevedocti@gmail.com

[1]Instituto de Ciências Matemáticas e Computação
Universidade de São Paulo, Brasil – São Paulo/SP

[2]Faculdade de Sistemas de informações
Instituto Federal do Tocantins - Paraíso do Tocantins/TO

[3]Divisão de Sistemas Ciberfisicos - DISCF
CTI/MCTI Renato Archer – Campinas/SP

Abstract. *This paper focuses on the use of emotion recognition techniques to assist psychologists in performing children's therapy through remotely robot operated sessions. In the field of psychology, the use of agent-mediated therapy is growing increasingly given recent advances in robotics and computer science. Specifically, the use of Embodied Conversational Agents (ECA) as an intermediary tool can help professionals connect with children who face social challenges such as Attention Deficit Hyperactivity Disorder (ADHD), Autism Spectrum Disorder (ASD) or even who are physically unavailable due to being in regions of armed conflict, natural disasters, or other circumstances. In this context, emotion recognition represents important feedback for the psychotherapist. In this article, we initially present the result of a bibliographical research associated with emotion recognition in children. This research revealed an initial overview on algorithms and datasets widely used by the community. Then, based on the analysis carried out on the results of the bibliographical research, we used the technique of dense optical flow features to improve the ability of identifying emotions in children in uncontrolled environments. From the output of a hybrid model of Convolutional Neural Network, two intermediary features are fused before being processed by a final classifier. The proposed architecture was called HybridCNNFusion. Finally, we present the initial results achieved in the recognition of children's emotions using a dataset of Brazilian children..*

Resumo. *Este trabalho foca no uso de técnicas de reconhecimento de emoções para auxiliar psicólogos na realização de terapia infantil por meio de sessões remotamente operadas por robôs. No campo da psicologia, o uso da terapia mediada por agentes está crescendo cada vez mais devido aos recentes avanços na robótica e na ciência da computação. Especificamente, o uso de Agentes de Conversação Incorporados (ECA em inglês) como uma ferramenta intermediária pode ajudar os profissionais a se conectarem com crianças que enfrentam desafios sociais como Transtorno de Déficit de Atenção e*

Hiperatividade (TDAH), Transtorno do Espectro Autista (TEA) ou mesmo que estão fisicamente indisponíveis devido a estarem em regiões de conflito armado, desastres naturais ou outras circunstâncias. Nesse contexto, o reconhecimento de emoções representa um feedback importante para o psicoterapeuta. Neste artigo, apresentamos inicialmente o resultado de uma pesquisa bibliográfica associada ao reconhecimento de emoções em crianças. Esta pesquisa revelou uma visão geral inicial sobre algoritmos e conjuntos de dados amplamente utilizados pela comunidade. Em seguida, com base na análise realizada sobre os resultados da pesquisa bibliográfica, utilizamos a técnica de características de fluxo óptico denso para melhorar a capacidade de identificar emoções em crianças em ambientes não controlados. A partir da saída de um modelo híbrido de Rede Neural Convolucional, duas características intermediárias são fundidas antes de serem processadas por um classificador final. A arquitetura proposta foi denominada HybridCNNFusion (fusão híbrida de Redes Neurais Convolucionais). Por fim, apresentamos os resultados iniciais alcançados no reconhecimento de emoções infantis utilizando um conjunto de dados de crianças brasileiras.

.

1. Introdução

O desenvolvimento cognitivo humano passa por várias etapas desde o nascimento até a maturidade. A infância representa a fase em que se adquire a base para aprender a se relacionar com os outros e com o mundo. Infelizmente, o processo de desenvolvimento mental de uma criança e a integração social como um todo pode ser prejudicada ou dificultada por transtornos mentais como ansiedade, estresse, comportamento obsessivo-compulsivo ou abuso emocional, sexual ou físico.

A solução ou redução das consequências dessas aflições é alcançada com processos terapêuticos realizados por profissionais da área da psicologia. Devido ao amadurecimento infantil limitado, o processo envolve não apenas sessões de avaliação com a criança, mas também entrevistas com pais e educadores, observação da criança nos ambientes residencial e escolar e coleta de dados por meio de desenhos, composições, jogos e outras atividades.

Nesse processo, os recursos de lazer como: jogos, atividades teatrais, fantoches, brinquedos e outros ganham destaque especial e são utilizados como apoio na terapia[21]. Como forma de contribuir com essa abordagem, Agentes Conversacionais Corporativos (ECA) são utilizados como ferramenta em aplicações psicoterapêuticas. Provoost et al. [23] realizaram uma scoping review sobre o uso de ‘ECAs’ em psicologia. Após a seleção, a busca revelou 49 trabalhos associados aos seguintes transtornos mentais: autismo, depressão, transtorno de ansiedade, transtorno de estresse pós-traumático, transtorno psicótico e uso de substâncias. Segundo os autores, "as aplicações de ‘ECAs’ são muito interessantes e apresentam resultados promissores, mas sua natureza complexa dificulta a comprovação de sua eficácia e segurança para uso na prática clínica". Na verdade, a estratégia sugerida por Provoost et al. envolve aumentar

a base de evidências por meio de intervenções usando agentes de baixa tecnologia que são rapidamente desenvolvidos, testados e aplicados na prática clínica responsável.

O reconhecimento das emoções durante as sessões psicoterapêuticas pode servir de auxílio ao profissional de psicologia envolvido no processo, havendo ainda um grande espaço para melhorias considerando a profundidade da tarefa em questão.

O objetivo deste trabalho é utilizar imagens geradas por câmeras presentes em uma sessão de psicoterapia de uma criança para classificar seu estado emocional em um determinado momento em uma das seguintes categorias básicas de emoção: raiva, nojo, medo, felicidade, tristeza, surpresa, desprezo [7]. Dada a diversidade de algoritmos de aprendizado de máquina para tarefas de reconhecimento de emoções em geral, abordar corretamente nosso objetivo é muito mais complexo do que simplesmente escolher o algoritmo mais poderoso ou recente [20]. Para aplicações em psicologia, em comparação com outras tarefas centradas no ser humano, a solução busca principalmente ser à prova de falhas e classificações incorretas e ser capaz de funcionar em cenários reais, o que é em si extremamente desafiador e, portanto, levanta múltiplas questões éticas e moralmente discutíveis sobre a viabilidade de tais modelos [14]. Nesse contexto, é importante estudar e considerar os ambientes para os quais um determinado algoritmo será utilizado antes mesmo de começar a desenvolvê-lo ou treiná-lo [19].

Na Seção [2], discutimos brevemente o estado atual da arte para tarefas de reconhecimento de emoções. Nossos conjuntos de dados de treinamento, bem como a arquitetura do modelo implementado, são apresentados nas Seções [3.1] e [3.2], respectivamente. Os resultados obtidos com o modelo sugerido e as conclusões do trabalho são discutidos nas Seções [4] e [5].

2. Estado da Arte

Uma pesquisa bibliográfica foi realizada para determinar o atual estado da arte (SOTA) para tarefas de reconhecimento de emoções usando algoritmos de computador. A busca foi feita no repositório ”Web of Science” [32], abrangendo os últimos 5 anos, com a seguinte chave de busca:

child AND emotion AND (recognition OR detection) AND*

(algorithm OR ”machine learning” OR ”computer vision”)

Um número inicial de 152 referências foi selecionado, com um total de 42 aceitos para leitura aprofundada (39 da busca original e 3 trabalhos extras). Uma análise de leitura adicional foi feita, marcando cada artigo de acordo com um número selecionado de categorias, incluindo, mas não se limitando a: conjuntos de dados usados; idade dos pacientes; procedimento psicológico adotado; formato de dados (como vídeo, fotos ou digitalizações); categoria de algoritmo (técnicas de aprendizado profundo, aprendizado de máquina puro, etc.). O resultado detalhado desta categorização pode ser visto na planilha disponível no Google Drive [1].

2.1 Tipos de algoritmos e conjuntos de dados

Na Figura [1] apresentamos os principais conjuntos de dados identificados durante a pesquisa bibliográfica. O conjunto de dados FER-2013 [24] é um dos mais utilizados pelos pesquisadores com 9 referências. Podemos citar os trabalhos de Sreedharan et al. [26] que fazem uso desse conjunto de dados, por exemplo.

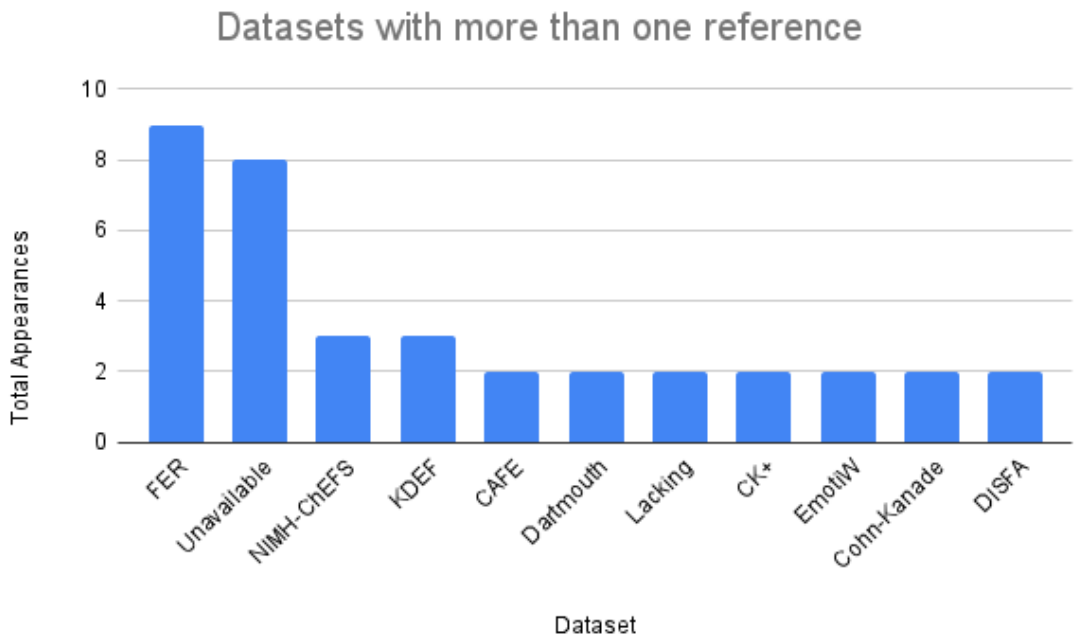

Figura 1: Conjuntos de dados usados para treinamento

No geral, descobrimos que os algoritmos de Reconhecimento de Emoção Facial (FER em inglês) tiveram melhorias significativas nos últimos anos [17], impulsionado pelo sucesso de abordagens baseadas em aprendizagem de redes profundas. Na Figura [2] apresentamos os algoritmos mais usados para reconhecimento de emoções. A arquitetura de redes neurais convolucionais (DL-CNN) foi a mais utilizada, com 22 referências. Como exemplos de DL-CNN, podemos citar os trabalhos de Haque e Valles [29] e Cuadrado et al. [13].

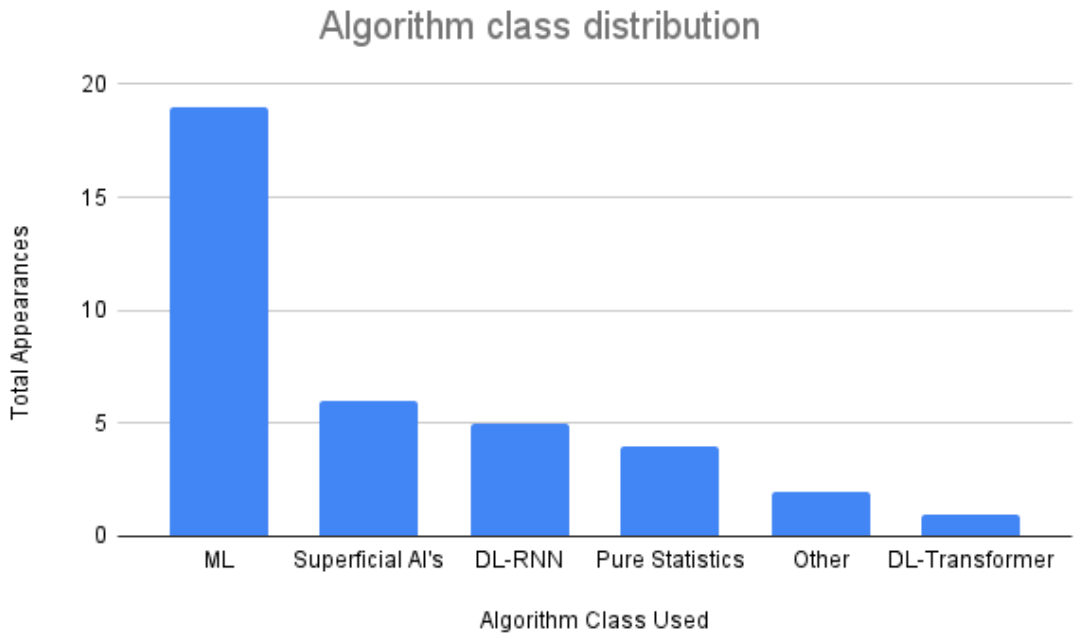

Figura 2. Algoritmos para reconhecimento de emoções.

Com a demanda por algoritmos de alto desempenho, vários novos modelos, como o sistema DeepFace para várias tarefas, como reconhecimento de emoções, regressão de idade, enquadramento de rostos [28] ou a arquitetura Transformer para sequências temporais de dados [30] também deram grandes passos para melhorar a precisão geral e a eficiência de tempo para modelos de classificação de emoção.

Entre os paradigmas mais populares atualmente usados para FER, as Redes Neurais Convolucionais (CNNs) demonstraram alto desempenho na detecção e reconhecimento de características emocionais de expressões faciais em imagens [14] aplicando filtros em movimento sobre uma imagem, também chamados de núcleos de convolução. Esses modelos usam técnicas de extração de recursos hierárquicos para construir informações baseadas em regiões a partir de imagens faciais, que são usadas para classificação. Um dos primeiros modelos amplamente difundidos baseados em CNN que foram usados para FER é a rede VGG-16, que usa 16 camadas de convolução e 3 camadas totalmente conectadas para classificar emoções [25]. Além das CNNs, outros modelos como Redes Neurais Recorrentes (RNNs) ou uma combinação de ambos também foram propostos para a tarefa de FER.

No geral, a área apresenta inovações recentes em pesquisa e há um trabalho contínuo para melhorar a precisão e a robustez das soluções existentes.

2.2 Estratégias clássicas de captura de emoções

Na Figura [3] apresentamos a origem das imagens estáticas presentes nos conjuntos de dados. Podemos observar que 48,8% dos estudos utilizaram emoções “Posadas”, de modo que as emoções expressas são artificiais, sendo sua promulgação solicitada por um avaliador. Como exemplo de trabalhos que usam emoções “Posadas” citamos Sreedharan et al. [29] e Goulart et al. [10]. O grupo de emoções “Induzidas” contribui com 23,3% dos trabalhos encontrados, para os quais podemos citar os trabalhos de Kahou et. al [17] e Kalantarian et. al [18]. O grupo “Espontâneo” aparece em apenas 16,3% dos estudos, possivelmente pela dificuldade em captar as emoções in-the-wild (ITW), ou seja, quando o indivíduo não tem conhecimento da finalidade ou da existência de gravação em vídeo em andamento ou fotografia.

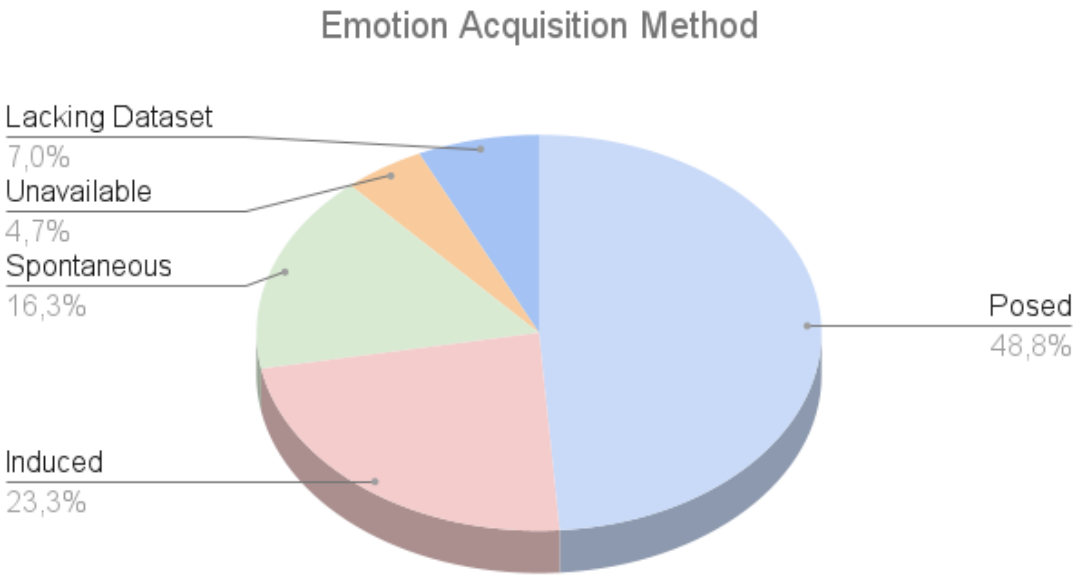

Figura 3: Estratégias de captura de emoções.

2.3 Arquiteturas Híbridas

Modelos que combinam várias redes em uma arquitetura, chamados de modelos híbridos, estão se tornando cada vez mais precisos, particularmente aqueles que combinam redes neurais convolucionais (CNNs) e redes neurais recorrentes (RNNs) [18] para tarefas de reconhecimento de emoções faciais (FER).

Além disso, pesquisas recentes mostraram que a integração de camadas recorrentes, como a camada de memória de longo e curto prazo (camada LSTM) [11], que processa entradas recursivamente, tornando-as particularmente úteis para capturar a dinâmica temporal de expressões faciais e inserir essas camadas em modelos híbridos podem melhorar ainda mais seu desempenho[12].

Outra linha de pesquisa promissora para melhorar a precisão na área de FER é o uso de vários recursos, como áudio e imagens processadas, além de imagens de cores faciais (RGB)[3], também chamado de aprendizado multimodal. Esses recursos adicionais podem fornecer informações complementares que podem melhorar a robustez e a precisão do sistema final. No entanto, ainda há desafios que precisam ser resolvidos, como fundir efetivamente vários recursos e como treinar efetivamente esses modelos em intervalos de tempo razoáveis. De qualquer forma, modelos híbridos que interpretem sequências de dados, como os *Transformers* ou classificadores com camadas LSTMs, bem como múltiplos recursos são uma direção promissora para melhorar o estado da arte em FER.

3. Metodologia

3.1. Conjuntos de dados de reconhecimento de emoções infantis

Considerando a necessidade de uma arquitetura que foque em dados não formatados (geralmente por serem ITW, ou seja, de crianças em ambientes não controlados) que também forneça precisão adequada e resposta em tempo real para prever emoções em crianças, criamos a arquitetura HybridCNNFusion para processar a sequência de quadros não formatados em tempo real.

Para realizar a tarefa em mãos, treinamos nosso modelo nos dois conjuntos de dados disponíveis publicamente com a maior precisão para tarefas FER [2]. Os conjuntos de dados usados são o FER-2013 [24] e o Karolinska Directed Emotional Faces [4]. Também ajustamos nosso modelo no conjunto de dados ChildEFES [22], um conjunto de dados privado, de vídeos e imagens contendo crianças brasileiras expondo alguma emoção dentro das 7 esperadas.

A maioria dos conjuntos de dados para tarefas FER é voltada para adultos e com expressões posadas, portanto, decidimos usar o conjunto de dados ChildEFES para ajuste fino (ou '*fine-tunning'* em inglês), bem como para testar o modelo final. Outros conjuntos de dados, como o Cohn-Kanade dataset (CK) e o Child Affective Facial Expression (CAFE) também foram considerados, mas apesar de públicos, seu uso tornou-se improvável devido à burocracia ou rejeição visto que nossos pedidos de acesso vieram de uma instituição localizada no brasil.

3.2. Modelo de Arquitetura HybridCNNFusion

Na Fig. 4 apresentamos os elementos que compõem a arquitetura HybridCNNFusion. A primeira etapa na construção de nossa arquitetura foi permitir que o modelo fosse usado em cenários selvagens, implementando um algoritmo de detecção de região Haarcascade [31] para centralizar e recortar os rostos das crianças.

Essas imagens cortadas são então passadas para uma Rede Neural Convolucional (CNN), especificamente a InceptionNet [27], para processar os pixels RGB cortados gerados pelo algoritmo Haarcascade. Paralelamente, usamos o algoritmo de fluxo óptico denso de Gunner-Farneback [9] para recuperar os valores de fluxo óptico denso dos quadros cortados atual e anterior. Isso é feito para permitir que a rede processe a variação nos músculos faciais e no movimento da pele ao longo do tempo.

As matrizes de fluxo óptico são então passadas para uma segunda CNN, especificamente uma variação do ResNet [10].

Depois de calcular esses dois conjuntos de valores processados separadamente, eles são concatenados e usados como entrada para um bloco recorrente final, feito especificamente com camadas de células LSTM para gerar a saída intermediária concatenada. Isso aproveita a natureza sequencial dos quadros de vídeo para produzir um vetor final de probabilidades previstas para cada emoção.

O modelo mencionado acima usa uma técnica chamada Late Fusion [12], na qual dois recursos separados são concatenados dentro da arquitetura e usados como entrada para as camadas de saída finais.

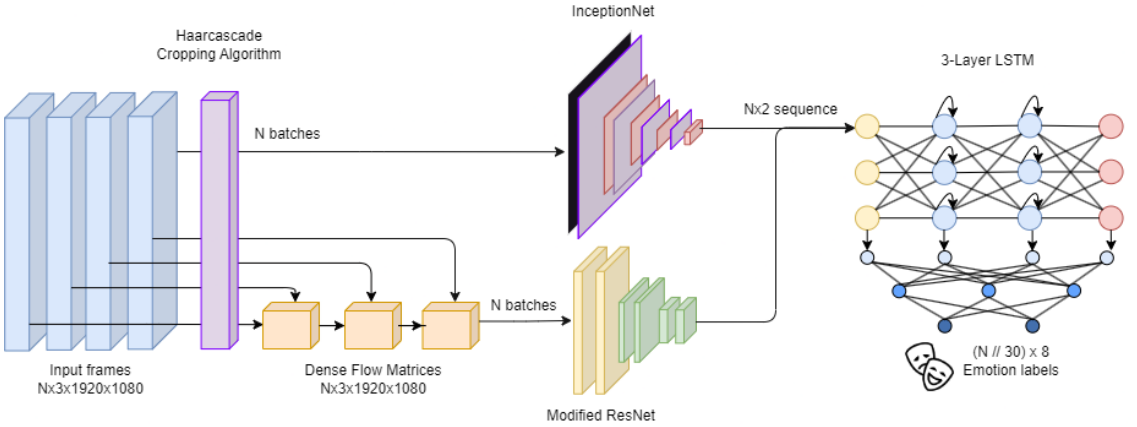

Figura 4. Arquitetura HybridCNNFusion. O código para esta arquitetura está disponível no GitHub

A técnica Late fusion permite um melhor uso do movimento gerado por Unidades de Ação faciais (*Facial Action Coding*) [6] por ter CNNs distintamente treinadas, uma para valores RGB brutos (saída pelo InceptionNet) e outra para matrizes de movimento HSV densas (saída pelo Combinação ResNet + OpticalFlow). O uso dos recursos de fluxo óptico como entrada para o ResNet permite o processamento de informações sequenciais, especificamente, de movimento, por meio da manipulação inteligente dos valores RGB brutos.

O passo a passo usado para uma única iteração de classificação de vídeo é apresentado na seção Algoritmo [1] abaixo.

Algoritmo 1 Pseudo-algoritmo do HybridCNNFusion

Entrada: Quadros RGB de tamanho $N \times 1920 \times 1080$ ($\vec{x}i$) e um vetor *one-hot* para o rótulo de emoção ao longo do vídeo (e).
Saída: Ej para cada janela de 10 segundos dos quadros.
Passos:
for $\vec{e_i}, i = 0 : M$ **do**
Recortar cada vetor $\vec{e_i}$ usando o algoritmo de recorte Haarcascade para centralizar os rostos,
para imagens de tamanho $n \times n$.
Aplicar o algoritmo de Gunner Farneback aos quadros $\vec{c_i}$ recortados.
Agrupar os recursos recortados e de fluxo óptico em grupos de 30 quadros.
Inserir ambos os grupos de entrada em duas redes CNN separadas, respectivamente:
CNNFlow = InceptionNet(3, 8) e CNNRaw = ResNet34(3, 8).
end for
for $group_j, 0 : N/30$ **do**
Concatenar os recursos recortados e de fluxo óptico.
Inserir os vetores concatenados em uma LSTM de 3 camadas e gerar uma sequência
de previsões baseadas nas probabilidades de emoção anteriores.
Anexar o rótulo de emoção do grupo a uma sequência de rótulos para o vídeo inteiro.
end for

3.3. Aspectos éticos e considerações da solução

A tarefa de reconhecimento de emoções faciais (FER) em crianças é particularmente desafiadora devido às questões éticas e à necessidade de alto nível de precisão e interpretabilidade.

A maioria das abordagens FER existentes concentra-se em situações não eticamente críticas, como a satisfação do cliente ou em condições de laboratório controladas[20]. Por outro lado, a tarefa de FER em crianças emocionalmente vulneráveis requer um nível muito maior de confiabilidade de acordo com as restrições éticas da relação psicólogo-paciente[5].

Este ramo de pesquisa específico das tarefas FER exige a capacidade de detectar e interpretar com precisão expressões faciais em vídeos em tempo real de crianças em situações selvagens (ITW), garantindo ao mesmo tempo a confiança das informações geradas [16] [19].

4. Resultados

O servidor utilizado para treinamento e teste da arquitetura possui limitações de memória e processamento que comprometem a implantação da arquitetura HybridCNNFusion. Nossa instituição deve habilitar o acesso a um servidor com mais recursos nas próximas semanas. A partir desse momento, espera-se que os testes de desempenho apresentem melhores resultados.

Apesar dessa limitação, o modelo final foi treinado nos conjuntos de dados FER2013 e KDEF e ajustado no conjunto de dados ChildEFES para maximizar a precisão. Devido a limitações de memória, o modelo inteiro não pôde ser totalmente ajustado em nosso conjunto de dados final, então medimos a precisão parcial para os modelos intermediários. O InceptionNet teve uma precisão de cerca de 70%, enquanto o ResNet teve uma precisão de cerca de 72%. No geral, o modelo teve uma única velocidade de iteração com média de 2,5s, para vídeos com média de 10s de duração.

As imagens de entrada são cortadas para o tamanho necessário de ambas as redes. A saída consiste em um vetor estocástico de probabilidades predizendo uma das 7 possíveis emoções básicas [8], bem como uma emoção neutra, totalizando 8 rótulos possíveis. Ambas as CNNs intermediárias possuem um vetor de saída de tamanho 32, e o recurso concatenado é um vetor com 64 entradas. A camada de saída final tem um tamanho de (N/ 30) × 8, com N/30 igual à duração total do vídeo dividido em grupos de 30 quadros, cada grupo com um rótulo de emoção previsto separado.

5. Conclusão

Considerando os aspectos tecnológicos e os resultados iniciais obtidos, a arquitetura proposta é um esforço contínuo para identificar as emoções das crianças em condições de natureza selvagem.

A fusão de recursos de fluxo óptico denso em conjunto com uma CNN híbrida e um modelo recorrente representa uma abordagem promissora na tarefa desafiadora A fusão de recursos de fluxo óptico denso em conjunto com uma CNN híbrida e um modelo recorrente representa uma abordagem promissora na tarefa desafiadora de reconhecimento de emoções faciais (FER) em crianças, especificamente em ambientes não controlados. Sendo uma necessidade crítica no campo da psicologia, esta abordagem oferece uma solução potencial.

Para situações eticamente sensatas, ainda existem métricas importantes que deveriam ser consideradas em cenários reais, como a Área sob a Curva ROC (AUC), que pode indicar se o modelo está propenso a perder emoções importantes em quadros pequenos e específicos , também chamadas de micro emoções [15].

De fato, existe uma grande lacuna nas questões éticas atuais para a tarefa, mas abordar essas métricas e melhorar a interpretabilidade, explicabilidade e segurança da transmissão da informação processada deve ser o foco de futuros modelos e frameworks. Isso garantirá que a tecnologia possa ser usada com segurança e eficácia para apoiar o bem-estar emocional das crianças.

6. Bibliografia

[1] Rafael Zimmer and Marcos Sobral and Helio Azevedo. Spreadsheet with Reference Classification Groups. bit.ly/3Oke2fR. accessed on 08 may 2023. 2023.

[2] De'Aira Bryant and Ayanna Howard. "A Comparative Analysis of Emotion Detecting Al Systems with Respect to Algorithm Performance and Dataset

Diversity". English. In: AIES '19: PROCEEDINGS OF THE 2019 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY. 1515 BROADWAY, NEW YORK, NY 10036-9998 USA: ASSOC COMPUTING MACHINERY, 2019, pp. 377–382. isbn: 978-1-4503-6324-2. doi: 10.1145/3306618. 3314284.

[3] M. Catalina Camacho, Helmet T. Karim, and Susan B. Perlman. "Neural architecture supporting active emotion processing in children: A multivariate approach". English. In: NEUROIMAGE 188 (Mar. 2019), pp. 171–180. issn: 1053-8119. doi: 10.1016/j.neuroimage.2018.12.013.

[4] A. Òhman D. E.Lundqvist A. Flykt. The Karolinska Directed Emotional Face. 1998. url: https://www.kdef.se/.

[5] Arnaud Dapogny et al. "On Automatically Assessing Children's Facial Expressions Quality: A Study, Database, and Protocol". English. In: FRONTIERS IN COMPUTER SCIENCE 1 (Oct. 2019). doi: 10.3389/fcomp.2019.00005.

[6] P. Ekman and W.V. Friesen. Facial Action Coding System. v. 1. Consulting Psychologists Press, 1978. url: https://books.google.com.br/books?id=08l6wgEACAAJ.

[7] Paul Ekman. "Are There Universal Facial Expressions? https://www.paulekman.com/resources/universal-facial-expressions/. Accessed on 10 fev 2023. 2022.

[8] Paul Ekman. "What Scientists Who Study Emotion Agree About". In: Perspectives on Psychological Science 11.1 (2016), pp. 31–34. doi: 10.1177/1745691615596992. eprint: https://doi.org/10.1177/1745691615596992. url: https://doi.org/10.1177/1745691615596992.

[9] Gunnar Farneback. Two-Frame Motion Estimation Based on Polynomial Expansion. Ed. by Josef Bigun and Tomas Gustavsson. Berlin, Heidelberg, 2003.

[10] Christiane Goulart et al. "Visual and Thermal Image Processing for Facial Specific Landmark Detection to Infer Emotions in a Child-Robot Interac- tion". English. In: SENSORS 19.13 (July 2019). issn: 1424-8220. doi: 10.3390/s19132844

[11] Kaiming He et al. Deep Residual Learning for Image Recognition. 2015. arXiv: 1512.03385 [cs.CV]

[12] Sepp Hochreiter and Jùrgen Schmidhuber. "Long short-term memory". In: Neural computation 9.8 (1997), pp. 1735–1780.

[13] Jiuk Hong, Chaehyeon Lee, and Heechul Jung. "Late Fusion-Based Video Transformer for Facial Micro-Expression Recognition". English. In: APPLIED SCIENCES-BASEL 12.3 (Feb. 2022). doi: 10.3390/app12031169.

[14] Luis-Eduardo Imbernon Cuadrado, Angeles Manjarres Riesco, and Felix de la Paz Lopez. “FER in Primary School Children for Affective Robot Tutors”. In: FROM BIOINSPIRED SYSTEMS AND BIOMEDICAL APPLICATIONS TO MACHINE LEARNING, PT II. Ed. by JMF Vicente et al. Vol. 11487. Lecture Notes in Computer Science. 8th International Work-Conference on the Interplay Between Natural and Artificial Computation (IWINAC), Almeria, SPAIN, JUN 03-07, 2019. Spanish CYTED; Red Nacl Computac Nat & Artificial, Programa Grupos Excelencia Fundac Seneca & Apliquem Microones 21 s l. GEWERBESTRASSE 11, CHAM, CH-6330, SWITZERLAND: SPRINGER INTERNATIONAL PUBLISHING AG, 2019, pp. 461–471. doi: 10.1007/978-3-030-19651-6 45.

[15] Asha Jaison and C. Deepa. “A Review on Facial Emotion Recognition and Classification Analysis with Deep Learning”. English. In: BIOSCIENCE BIOTECHNOLOGY RESEARCH COMMUNICATIONS 14.5, SI (2021), pp. 154–161. issn: 0974-6455. doi: 10.21786/bbrc/14.5/29.7

[16] Salma Kammoun Jarraya, Marwa Masmoudi, and Mohamed Hammami. “Compound Emotion Recognition of Autistic Children During Meltdown Crisis Based on Deep Spatio-Temporal Analysis of Facial Geometric Features”. English. In: IEEE ACCESS 8 (2020), pp. 69311–69326. issn: 21693536. doi: 10.1109/ACCESS.2020.2986654.

Haik Kalantarian et al. “The Performance of Emotion Classifiers for Children With Parent-Reported Autism: Quantitative Feasibility Study”. English. In: JMIR MENTAL HEALTH 7.4 (Apr. 2020). issn: 2368-7959. doi:10.2196/13174.

[17] Akhilesh Kumar and Awadhesh Kumar. “Analysis of Machine Learning Algorithms for Facial Expression Recognition”. English. In: ADVANCED NETWORK TECHNOLOGIES AND INTELLIGENT COMPUTING, ANTIC 2021. Vol. 1534. 2022, pp. 730–750. isbn: 978-3-030-96040-7; 978-3-030-96039-1. doi: 10.1007/978-3-030-96040-7_55.

[18] Sunan Li et al. “Bi-modality Fusion for Emotion Recognition in the Wild”. English. In: ICMI 19: PROCEEDINGS OF THE 2019 INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION. 21st ACM International Conference on Multimodal Interaction (ICMI), Suzhou, PEOPLES R CHINA, OCT 14-18, 2019. ACM SIGCHI; Assoc Comp Machinery; Openstream; Alibaba Grp; Microsoft; Baidu; Sensetime; Tencent YouTu Lab; AISpeech. 1601 Broadway, 10th Floor, NEW YORK, NY, UNITED STATES: ASSOC COMPUTING MACHINERY, 2019, pp. 589–594. isbn: 978-1-4503-6860-5. doi: 10.1145/3340555.3355719.

[19] Jose Luis Espinosa-Aranda et al. “Smart Doll: Emotion Recognition Using Embedded Deep Learning”. English. In: SYMMETRY-BASEL 10.9 (Sept.2018). issn: 2073-8994. doi: 10.3390/sym10090387.

[20] Aleix M. Martinez. “The Promises and Perils of Automated Facial Action Coding in Studying Children’s Emotions”. English. In: DEVELOPMENTAL PSYCHOLOGY 55.9, SI (Sept. 2019), pp. 1965–1981. issn: 0012-1649. doi: 10.1037/dev0000728.

[21] Cynthia Borges de Moura and M.R.Z.S. Azevedo. “Estrategias lúdicas para uso em terapia comportamental infantil”. In: Sobre comportamento e cognição: questionando e ampliando a teoria e as intervenções clínicas e em outros contextos. Ed. by R. C. Wielenska. Vol. 6. Santo André, 2000, pp. 163–170.

[22] Juliana Gioia Negrão et al. “The Child Emotion Facial Expression Set: A Database for Emotion Recognition in Children”. In: Frontiers in Psychology 12 (2021). issn: 1664-1078. doi: 10.3389/fpsyg.2021.666245. url: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.666245.

[23] Simon Provoost et al. “Embodied Conversational Agents in Clinical Psychology: A Scoping Review”. In: Journal of Medical Internet Research 19.5 (May 2017), e151. issn: 1438-8871. doi: 10.2196/jmir.6553.

[24] Manas Sambare. ”FER-2013 Learn facial expressions from an image. https://www.kaggle.com/datasets/msambare/fer2013. accessed on 15 fev 2023. 2022.

[25] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2015. arXiv: 1409.1556 [cs.CV].

[26] Ninu Preetha Nirmala Sreedharan et al. “Grey Wolf optimisation-based feature selection and classification for facial emotion recognition”. In: IET BIOMETRICS 7.5 (Sept. 2018), pp. 490–499. issn: 2047-4938. doi: 10.1049/iet-bmt.2017.0160.

[27] Christian Szegedy et al. Going Deeper with Convolutions. 2014. arXiv:1409.4842 [cs.CV].

[28] Yaniv Taigman et al. “DeepFace: Closing the Gap to Human-Level Performance in Face Verification”. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, June 2014. doi: 10.1109/cvpr.2014.220. Hybrid Models for Facial Emotion Recognition in Children 11.

[29] Md Inzamam Ul Haque and Damian Valles. “A Facial Expression Recognition Approach Using DCNN for Autistic Children to Identify Emotions”. In: 2018 IEEE 9TH ANNUAL INFORMATION TECHNOLOGY, ELECTRONICS AND MOBILE COMMUNICATION CONFERENCE (IEMCON). Ed. by S Chakrabarti and HN Saha. 9th IEEE Annual Information Technology, Electronics and Mobile Communication Conference

(IEMCON), Univ British Columbia, Vancouver, CANADA, NOV 01-03, 2018. Inst Engn & Management; IEEE Vancouver Sect; UBC; Univ Engn

& Management. IEEE, 2018, pp. 546–551.

[30] Ashish Vaswani et al. Attention Is All You Need. 2017. arXiv: 1706.03762 [cs.CL].

[31] Paul Viola and Michael Jones. “Rapid Object Detection using a Boosted Cascade of Simple Features”. In: CONFERENCE ON COMPUTER VI-SION AND PATTERN RECOGNITION (2001).

[32] Web of Science. ”Web of Science platform. bit.ly/3McZko4. accessed on

08 may 2022. 2022

[33] Kaiming He et al. Deep Residual Learning for Image Recognition. 2015. arXiv: 1512.03385 [cs.CV]

[34] Samira Ebrahimi Kahou et al. “Recurrent Neural Networks for Emotion Recognition in Video”. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, Nov. 2015. doi: 10.1145/ 2818346.2830596

[35] Haik Kalantarian et al. “The Performance of Emotion Classifiers for Children With Parent-Reported Autism: Quantitative Feasibility Study”. English. In: JMIR MENTAL HEALTH 7.4 (Apr. 2020). issn: 2368-7959. doi: 10.2196/13174